Machine Learning Engineer Nanodegree

Capstone Proposal

Stuart McMeechan
February 2018

Proposal

Domain Background

Determining the health of a business, or an entity within a business, is an important activity for a number of business reasons. For example, analysis of the strength of a business, focussed particularly the balance sheet, will be carried out as standard practise when merging with or acquiring a company. Other examples of when this would be carried out could be during the onboarding of a new supplier or customer, or when carrying out an internal risk review of entities across a global business.

An example of the importance of effectively screening new suppliers relates to the recent collapse of one of the UK's largest construction services companies. When the company entered administration in 2018, it had £1.5bn of debt and owed up to 30,000 businesses approximately £800m in payments. Could these businesses have been aware of the impending collapse of the company, which could either have been their supplier (e.g. playing a key part in their construction project) or customer (e.g. buying parts or people from them)?

Problem Statement

Businesses go bankrupt and enter administration regularly. The problem statement is, given a limited set of financial data typically publicly available, can you predict if a business is in financial distress and likely to collapse?

It is likely that rule based analysis is common, for example placing a high risk factor when the business being analysed has been selling a high percentage of their fixed assets or if their cash flow is under a specified threshold. The problem statement is, can a machine learning algorithm be used to more accurately forecast whether a company is likely to go into bankruptcy or not.

Datasets and Inputs

A dataset related to companies going bankrupt was found on the UCI Machine Learning Repository, via Kaggle. The dataset contains five files, reflecting the financial metrics of companies from 2007-2013:

1. Year 1: Financial metrics for companies, with a tag showing if they were bankrupt 5 years after this time.

2. Year 2: Financial metrics for companies, with a tag showing if they were bankrupt 4 years after this time.
3. Year 3: Financial metrics for companies, with a tag showing if they were bankrupt 3 years after this time.
4. Year 4: Financial metrics for companies, with a tag showing if they were bankrupt 2 years after this time.
5. Year 5: Financial metrics for companies, with a tag showing if they were bankrupt 1 year after this time.

There are 64 features in each file and most of these are calculated values, such as net profit divided by total assets and net profit divided by inventory, all using financial metrics found on a company balance sheet. Companies cannot be identified across datasets, so it is only possible to view the financial metrics for company on a given year.

Since the dataset includes a tag when any of the companies went bankrupt, it can be used to assess whether any of the financial metrics show a difference between companies that we know go bankrupt and those that don't.

https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data

Solution Statement

The business question is, given a historic set of financial metrics about a company, can you accurately predict whether it will go bankrupt in the near future? The solution to this will involve building and testing a supervised learning model, using the dataset described in the previous section.

I propose that this is a multiclass classification problem, with three possible categories for each company: Predicted to survive (0), Predicted to go bankrupt within the next 2 years (1) and Predicted to go bankrupt in 3-5 years (2). I propose to test the following four types of supervised learning algorithms on the dataset:

- Neural Network
- Support Vector Machine
- Decision Tree
- K-nearest Neighbour

At the end of the analysis, I plan to have the following:

- A view on which features are best for training the algorithm
- A view on which type of algorithm performs best on the testing data
- A trained algorithm, with an evaluation score, which can be used to predict whether companies are likely to go bankrupt

Benchmark Model

A journal published on PLOS in 2016 shows analysis of key financial metrics to predict if a company is likely to enter bankruptcy or not. The model's performance, split by the region of each company tested, has a top accuracy score of 90%. I will use this to benchmark my models against.

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0166693#sec011

Evaluation Metrics

I will use Accuracy ((tp+tn) / (tp+tn+fp+fn)) and F-score (formula below) to evaluate the models tested.

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

The F-score, which considers both precision and recall, is required as an evaluation metric since the classification distribution is skewed; only 3-4% of companies in the dataset go bankrupt.

I plan to evaluate the four types of algorithms listed in the previous section and select the one with the highest performance.

Project Design

Analysis Environment
I will use Python 2 for the core analysis and leverage the sklearn package for model build and testing. I may use SQL and Tableau for any initial data profiling and feature transformation.

Analysis Lifecycle
I plan to follow the following process for the capstone project.

1. Data loading and cleanup
The source CSV files will be loaded and unioned into a single dataframe. I will then transform the 'bankruptcy' column so it aligns to the target classifications, e.g.:

- Value 0 will be assigned to companies that didn't go bankrupt
- Value 1 will be assigned for companies tagged in files 4-5 (as they go bankrupt within 1-2 years)
- Value 2 will be assigned for companies tagged in files 1-3 (as they go bankrupt between three and five years later)

Initial profiling will then be carried out to answer questions such as:

- How many companies are in the dataset

- What percentage of the companies go bankrupt
- What percentage of the companies go bankrupt, split by classification value

2. Splitting data into training and testing

The dataset will be split using sklearn's cross validation function and the training dataset will be used until the model evaluation stage. *80% of the data will be used for training and 20% for testing.*

I will carry out some initial profiling of the training dataset to ensure there are sufficient examples of each target classification.

3. Feature selection / transformation

To start this stage I plan to:

- Transform any skewed continuous features, identified at the profiling stage, so that any very large or small values do not negatively affect the algorithm performance; and
- Normalise any numerical features to ensure each feature is treated equally.

I then plan to extract feature importances using sklearn's feature importance function. This will provide some initial guidance on feature selection across all features and I may remove some features if they don't have enough impact.

4. Model training and evaluation

The final stage will involve training the following types of algorithms:

- K-nearest Neighbour
- Support Vector Machine
- Neural Network
- Decision Tree

For each algorithm I will use sklearn to calculate the accuracy and F-score. Depending on the results, I may re-evaluate the feature selection and transformation stage. After achieving the best performing algorithm I can, ideally one which is close to the results I am benchmarking against, I will conclude the analysis.

Appendix A - Training Data Features

X1 net profit / total assets
X2 total liabilities / total assets
X3 working capital / total assets
X4 current assets / short-term liabilities
X5 [(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365
X6 retained earnings / total assets
X7 EBIT / total assets
X8 book value of equity / total liabilities
X9 sales / total assets
X10 equity / total assets
X11 (gross profit + extraordinary items + financial expenses) / total assets
X12 gross profit / short-term liabilities
X13 (gross profit + depreciation) / sales
X14 (gross profit + interest) / total assets
X15 (total liabilities * 365) / (gross profit + depreciation)
X16 (gross profit + depreciation) / total liabilities
X17 total assets / total liabilities
X18 gross profit / total assets
X19 gross profit / sales
X20 (inventory * 365) / sales
X21 sales (n) / sales (n-1)
X22 profit on operating activities / total assets
X23 net profit / sales
X24 gross profit (in 3 years) / total assets
X25 (equity - share capital) / total assets
X26 (net profit + depreciation) / total liabilities
X27 profit on operating activities / financial expenses
X28 working capital / fixed assets
X29 logarithm of total assets
X30 (total liabilities - cash) / sales
X31 (gross profit + interest) / sales
X32 (current liabilities * 365) / cost of products sold
X33 operating expenses / short-term liabilities
X34 operating expenses / total liabilities
X35 profit on sales / total assets
X36 total sales / total assets
X37 (current assets - inventories) / long-term liabilities
X38 constant capital / total assets
X39 profit on sales / sales
X40 (current assets - inventory - receivables) / short-term liabilities
X41 total liabilities / ((profit on operating activities + depreciation) * (12/365))
X42 profit on operating activities / sales
X43 rotation receivables + inventory turnover in days

X44	(receivables * 365) / sales
X45	net profit / inventory
X46	(current assets - inventory) / short-term liabilities
X47	(inventory * 365) / cost of products sold
X48	EBITDA (profit on operating activities - depreciation) / total assets
X49	EBITDA (profit on operating activities - depreciation) / sales
X50	current assets / total liabilities
X51	short-term liabilities / total assets
X52	(short-term liabilities * 365) / cost of products sold)
X53	equity / fixed assets
X54	constant capital / fixed assets
X55	working capital
X56	(sales - cost of products sold) / sales
X57	(current assets - inventory - short-term liabilities) / (sales - gross profit - depreciation)
X58	total costs /total sales
X59	long-term liabilities / equity
X60	sales / inventory
X61	sales / receivables
X62	(short-term liabilities *365) / sales
X63	sales / short-term liabilities
X64	sales / fixed assets